Communications of the Korean Statistical Society
2012, Vol. 19, No. 5, 743–750

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.5.743

Concept of the One-Sided Variance with Applications

Hyo-Il Park[1,a]

[a]Department of Statistics, Chongju University

Abstract

In this study, we propose definitions for the one-sided variance for asymmetric distribution. We consider to apply the one-sided variance to the construction to define modified C_{pk}, which is a definition for the process capability index for the asymmetric process distribution. Then we consider to obtain the consistent estimation for the one-sided variance and to apply to the various industrial fields.

Keywords: Asymmetry, control chart, process capability index.

1. Introduction

Let X be a random variable having an unknown but continuous distribution function F with finite second moment. Then the mean μ and variance σ^2 are defined as

$$\mu = E(X) = \int_{-\infty}^{\infty} x\, dF(x) \quad \text{and} \quad \sigma^2 = E\left\{(X-\mu)^2\right\} = \int_{-\infty}^{\infty} (x-\mu)^2 dF(x).$$

We note that the variance σ^2 is defined as the mean of the square of the deviation from the mean μ. Therefore, σ^2 is an average for the square of the deviation for both sides around the mean μ. Then if F is symmetric, this definition for σ^2 reveals no problem when we define any concept that should be represented by limits for both sides around μ based on σ^2. However for the non-symmetric or asymmetric case, the definition for σ^2 may incur some inconvenient or absurd situations. For example, one may consider to apply the process capability index(PCI) to assess the state of the ability for the production process and use the following C_{pk} among the various definitions for the PCI when the process distribution F is non-symmetric.

$$C_{pk} = \min\left\{\frac{\mu - \text{LSL}}{3\sigma}, \frac{\text{USL} - \mu}{3\sigma}\right\}, \tag{1.1}$$

where LSL and USL are the lower and upper specification limits, respectively. To get some more specific insight for the motivation of this study, we consider to provide the numerical values of C_{pk} for the Weibull distributions, that consist of a family for the skewed distributions. The probability density function(pdf) for any Weibull distribution has the following form: for any $\alpha > 0$,

$$f(x) = \begin{cases} \alpha x^{\alpha-1} \exp\left[-x^{\alpha}\right], & x > 0, \\ 0, & x \le 0. \end{cases}$$

In Table 1, we tabulated the values of C_{pk} by varying the value of α. We considered three cases such as 1/2, 1 and 2 for the values of α. In addition, we considered LSL = $w_{0.005}$ and USL = $w_{0.995}$, where w_p means the p^{th} quantile point for each Weibull distribution.

[1] Professor, Department of Statistics, Chongju University, Chongju 360-764, Korea, E-mail: hipark@cju.ac.kr

Table 1: Values of C_{pk} based on σ^2

α	mean	variance	LSL	USL	$(\mu - \text{LSL})/3\sigma$	$(\text{USL} - \mu)/3\sigma$	C_{pk}
1/2	2	20	0.0000	28.0722	0.1491	1.9433	0.1491
1	1	1	0.0050	5.2983	0.3317	1.4328	0.3317
2	$\sqrt{\pi}/2$	$1 - \pi/4$	0.0708	2.3018	0.5867	1.0186	0.5867

Table 2: A criterion for the assessment of process for C_p

Range of C_p	$C_p \geq 1.33$	$1 \leq C_p < 1.33$	$0.67 \leq C_p < 1$	$C_p < 0.67$
Grade of C_p	*A*	*B*	*C*	*D*

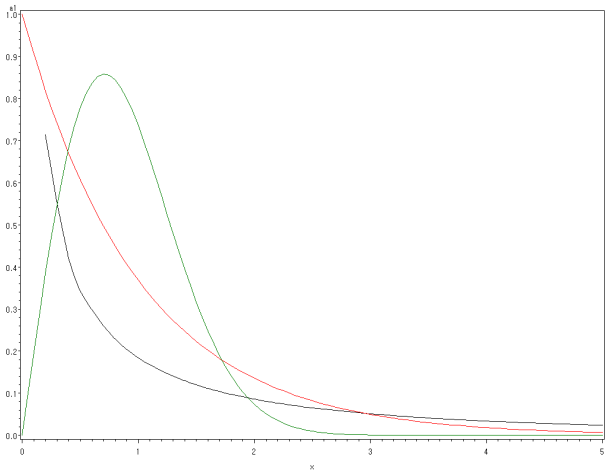

Figure 1: *The shape of pdf of Weibull distributions*

¿From Table 1, we note that the value of C_{pk} varies with a wide range(from 0.1491 to 0.5867) by the value of α. Therefore it would be difficult to establish an effective criterion for C_{pk} to assess the production process unlike C_p, which is also a PCI used mainly under the assumption of the normality for the process distribution. We note that C_p has a general criterion, which can play a useful role in monitoring the ability of the production process, with some specific values for the assessment of the ability of the process (Park and Park, 2005). In Table 2, we summarized such a criterion for C_p. Then the wide range of the value of C_{pk} in Table 1 may be because the shape of the Weibull distribution changes abruptly with the change of the value of as shown in Figure 1. We note that as α increases, the shape of the pdf moves toward the symmetry and the difference between $(\mu - \text{LSL})/3\sigma$ and $(\text{USL} - \mu)/3\sigma$ decreases. From this, one may conjecture that this phenomenon may happen since the shape of the pdf changes toward symmetry.

We note that the symmetry of F means that the deviations for the left and right sides from μ are equal. However, they are different when F is not symmetric. Therefore, since C_{pk} has been defined with the ordinary variance σ^2, it seems that the variation with such a wide range for the values of C_{pk} shown in Table 1 would be inevitable. Therefore in order to prevent or at least alleviate the excessive variation for the value of C_{pk} from σ^2 and help set up any criterion to easily monitor the production process, we will introduce a new definition for the variance that may measure the deviation with the sidewise manner for any given point. In the next section, we propose definitions for the one-sided variance by defining the modified C_{pk}s and comparing C_{pk} with the modified C_{pk}s by obtaining the

Table 3: The values of C_{pk}^{m1} based on $\sigma_L^2(\mu)$ and $\sigma_R^2(\mu)$

α	mean	$\sigma_L^2(\mu)$	$\sigma_R^2(\mu)$	LSL	USL	$\frac{\mu - \text{LSL}}{3\sigma_L}$	$\frac{\text{USL} - \mu}{3\sigma_R}$	C_{pk}^{m1}
1/2	2	2.6737	$40 + 24\sqrt{2}$	0.0000	28.0722	0.5130	1.0131	0.5130
1	1	0.4180	2	0.0050	5.2983	0.3317	1.4328	0.3317
2	$\sqrt{\pi}/2$	0.1630	0.2762	0.0708	2.3018	0.6733	0.8979	0.6733

values from the Weibull distributions used for Table 1. In Section 3, we consider the estimation and discuss some interesting features.

2. Definitions of the One-Sided Variance and Applications

In this section we define two types of the one-sided variance in the sequel. First of all, we propose the first definition for the one-sided variance from any given point $u \in (-\infty, \infty)$ as follows:

$$\sigma_L^2(u) = \int_{-\infty}^{u} (x-u)^2 d\frac{F(x)}{F(u)} = \frac{1}{F(u)} \int_{-\infty}^{u} (x-u)^2 dF(x)$$

and

$$\sigma_R^2(u) = \int_{u}^{\infty} (x-u)^2 d\frac{F(x)}{1-F(u)} = \frac{1}{1-F(u)} \int_{u}^{\infty} (x-u)^2 dF(x).$$

We note that $\sigma_L^2(u)$ and $\sigma_R^2(u)$ measure scales for the left and right side from u and may be called the left- and right-side variances from u, respectively. One may choose the point u according to the purpose of the application. For the application to C_{pk}, one should take $u = \mu$. If one considers a control chart based on a median, one has to choose a median for u. We note that if F is symmetric with mean μ, then we see that

$$\sigma_L^2(\mu) = \sigma_R^2(\mu) = \sigma^2.$$

Thus the one-sided variances $\sigma_L^2(u)$ and $\sigma_R^2(u)$ can be considered as a generalization of σ^2. Based on this definition for the one-sided variances, one may propose a modified definition C_{pk}^{m1} of C_{pk} as follows:

$$C_{pk}^{m1} = \min\left\{\frac{\mu - \text{LSL}}{3\sigma_L(\mu)}, \frac{\text{USL} - \mu}{3\sigma_R(\mu)}\right\}.$$

In Table 3, we summarized the values of C_{pk}^{m1} for the three Weibull distributions considered in Section 1 to compare with those of C_{pk}. We note that the variability of the value of C_{pk}^{m1} is considerably mitigated compared with that of C_{pk}. Therefore one may provide a sensible guideline to assess the production process with some common criterion at least within this Weibull distribution family.

Another concept for the one-sided variance can be defined as follows. For this, for any given point $u \in (-\infty, \infty)$, let

$$\mu_L(u) = \frac{1}{F(u)} \int_{-\infty}^{u} x dF(x) \quad \text{and} \quad \mu_R(u) = \frac{1}{1-F(u)} \int_{u}^{\infty} x dF(x).$$

We note that when $u = \infty$ or $u = -\infty$, we see that

$$\mu_L(\infty) = \mu_R(-\infty) = \mu.$$

Table 4: The values of C_{pk}^{m2} based on $\sigma_L^2(\mu_L)$ and $\sigma_R^2(\mu_R)$

α	mean	$\sigma_L^2(\mu_L)$	$\sigma_R^2(\mu_R)$	LSL	USL	$\frac{\mu - \text{LSL}}{3\sigma_L}$	$\frac{\text{USL} - \mu}{3\sigma_R}$	C_{pk}^{m2}
1/2	2	0.2683	$28 + 16\sqrt{2}$	0.0000	28.0722	1.2870	1.2214	1.2214
1	1	0.0793	1	0.0050	5.2983	1.1776	1.4328	1.1776
2	$\sqrt{\pi}/2$	0.0459	0.1094	0.0708	2.3018	1.2691	1.4263	1.2691

Thus μ_L and μ_R can be considered as the means of the truncated distributions from right and left at u. Then the one-sided variances for the left and right side, $\sigma^2(\mu_L(u))$ and $\sigma^2(\mu_R(u))$ can be defined as

$$\sigma^2(\mu_L(u)) = \int_{-\infty}^{u} (x - \mu_L(u))^2 d\frac{F(x)}{F(u)} = \frac{1}{F(u)} \int_{-\infty}^{u} (x - \mu_L(u))^2 dF(x)$$

and

$$\sigma^2(\mu_R(u)) = \int_{u}^{\infty} (x - \mu_R(u))^2 d\frac{F(x)}{1 - F(u)} = \frac{1}{1 - F(u)} \int_{u}^{\infty} (x - \mu_R(u))^2 dF(x).$$

Then we note that if $u = \infty$ or $u = -\infty$, since

$$\sigma^2(\mu_L(u)) = \sigma^2(\mu_R(u)) = \sigma^2$$

one may consider $\sigma^2(\mu_L(u))$ and $\sigma^2(\mu_R(u))$ generalizations of σ^2. With this definition for the one-sided variance, we may propose another modified definition C_{pk}^{m2} for C_{pk} by taking $u = \mu$ as follows.

$$C_{pk}^{m2} = \min\left\{\frac{\mu - \text{LSL}}{3\sigma_L(\mu_L)}, \frac{\text{USL} - \mu}{3\sigma_R(\mu_R)}\right\}.$$

Using this definition for the one-sided variance, we have summarized the values of C_{pk}^{m2} for the Weibull distributions considered in the previous section in Table 4. Then we note that the values of C_{pk}^{m2} maintain quite stable phase within the Weibull distribution family. In addition, we note that the two components comprised C_{pk}^{m2} show very little difference for each distribution and furthermore the variability of C_{pk}^{m2} becomes very stabilized. Therefore it would be possible to set up a guideline for PCI to easily control the production process.

3. Estimation of the One-Sided Variance

For this let $X_1, \ldots, X_n$ be a sample from a production process having the unknown but continuous distribution function F with the finite variance σ^2. First, we consider the estimation of σ_L^2 and σ_R^2. For this let $\hat{F}_n$ be the empirical distribution function from $X_1, \ldots, X_n$ such as

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \le x),$$

where $I(\cdot)$ is an indicator function. Then for any given real number $u \in (-\infty, \infty)$, estimates $\hat{\sigma}_L^2$ and $\hat{\sigma}_R^2$ of σ_L^2 and σ_R^2 can be proposed as follows:

$$\hat{\sigma}_L^2(u) = \frac{1}{\hat{F}_n(u)} \int_{-\infty}^{u} (x - u)^2 d\hat{F}_n(x) = \frac{1}{n\hat{F}_n(u)} \sum_{i=1}^{n} (X_i - u)^2 I(X_i \le u)$$

and

$$\hat{\sigma}_R^2(u) = \frac{1}{1-\hat{F}_n(u)} \int_u^{\infty} (x-u)^2 d\hat{F}_n(x) = \frac{1}{n\left(1-\hat{F}_n(u)\right)} \sum_{i=1}^{n} (X_i-u)^2 I(X_i > u).$$

Theorem 1. *With the condition of the finiteness for σ^2, $\hat{\sigma}_L^2$ and $\hat{\sigma}_R^2$ are the consistent estimates of σ_L^2 and σ_R^2.*

Proof: We only prove the consistency for $\hat{\sigma}_L^2$. The proof for $\hat{\sigma}_R^2$ follows the same arguments used for $\hat{\sigma}_L^2$. Then first of all, we note that

$$E\left\{(X_i-u)^2 I(X_i \le u)\right\} = \int_{-\infty}^{u} (x-u)^2 dF(x) < \infty.$$

Thus due to the Khintchine's weak convergence (Chung, 1974), we see that

$$\frac{1}{n}\sum_{i=1}^{n} (X_i-u)^2 I(X_i \le u) \xrightarrow{p} \int_{-\infty}^{u} (x-u)^2 dF(x),$$

where $\xrightarrow{p}$ means the convergence in probability. In addition, it is well-known that from the law of large numbers for every $u \in (-\infty, \infty)$,

$$\hat{F}_n(u) \xrightarrow{p} F(u).$$

Then from the Slutsky's theorem (Bickel and Doksum, 1977), we obtain that for any given real number $u \in (-\infty, \infty)$

$$\hat{\sigma}_L^2(u) \xrightarrow{p} \sigma_L^2(u).$$

For the applications to the real situation, u may be the mean of F or a median. In general, since they are unknown, one should estimate u to complete an estimate for the one-sided variance. Then for any consistent estimate $\hat{u}$ of u, consistent estimates $\hat{\sigma}_L^2$ and $\hat{\sigma}_R^2$ can be proposed as follows:

$$\hat{\sigma}_L^2(\hat{u}) = \frac{1}{n\hat{F}_n(\hat{u})} \sum_{i=1}^{n} (X_i-\hat{u})^2 I(X_i \le \hat{u}) \quad \text{and} \quad \hat{\sigma}_R^2(\hat{u}) = \frac{1}{n\left(1-\hat{F}_n(\hat{u})\right)} \sum_{i=1}^{n} (X_i-\hat{u})^2 I(X_i > \hat{u}).$$

The proof for the consistency would be straightforward if we apply the same arguments used for the proof of Theorem 1. For the consistent estimates of $\sigma^2(\mu_L(u))$ and $\sigma^2(\mu_R(u))$, first, we have to estimate $\mu_L(u)$ and $\mu_R(u)$. For this, let $\hat{u}$ be any consistent estimate of u. Then consistent estimates of $\mu_L(u)$ and $\mu_R(u)$ can be proposed as

$$\hat{\mu}_L(\hat{u}) = \frac{1}{n\hat{F}_n(\hat{u})} \sum_{i=1}^{n} X_i I(X_i \le \hat{u}) \quad \text{and} \quad \hat{\mu}_R(\hat{u}) = \frac{1}{n\left(1-\hat{F}_n(\hat{u})\right)} \sum_{i=1}^{n} X_i I(X_i > \hat{u}).$$

Then we may propose consistent estimate for $\sigma^2(\mu_L(u))$ and $\sigma^2(\mu_R(u))$ as follows:

$$\hat{\sigma}(\hat{\mu}_L(\hat{u})) = \frac{1}{n\hat{F}_n(\hat{u})} \sum_{i=1}^{n} (X_i-\hat{\mu}_L(\hat{u}))^2 I(X_i \le \hat{u})$$

and

$$\hat{\sigma}\left(\hat{\mu}_R(\hat{u})\right) = \frac{1}{n\left(1-\hat{F}_n(\hat{u})\right)} \sum_{i=1}^{n} \left(X_i - \hat{\mu}_R(\hat{u})\right)^2 I(X_i > \hat{u}).$$

The proof for the consistency would be straightforward if one uses the same arguments with the proof of Theorem 1. □

4. More Applications and Some Concluding Remarks

We have already applied the concept of the one-sided variance to define new PCIs for the asymmetric process distributions. The main purpose of the introduction of this concept was to be able to set up some useful criteria to grasp easily and quickly the situation of the production process with grading the modified C_{pk} that uses one of the proposed one-sided variances as we may have done for the case of C_p in Table 2. Then based on this point of view, the second definition would be more appropriate than the first one from the results of Table 3 and Table 4 since the values of C_{pk}^{m2} appear more evenly than those of C_{pk}^{m1}. The reasons for this are unclear; however, it may be because the one-sided variance from the second definition can be considered as the true variance for that side while that from the first one, as the variance of the symmetric distribution centered at u for each side.

As another application, one may consider to use to construct the control charts for the asymmetric underlying distributions. As an example, we consider the Weibull distribution with $\alpha = 1$ introduced in Section 1 with the sample size $n = 9$. Then from Table 1, the lower control limit(LSL) and upper control limit(USL) may be chosen as

$$\text{LCL} = \mu - \frac{3\sigma}{\sqrt{n}} = 1 - 1 = 0,$$
$$\text{UCL} = \mu + \frac{3\sigma}{\sqrt{n}} = 1 + 1 = 2,$$

when we consider $3-\sigma$ control limits. Then we note that we come to fail to control the lower part of the quality of product since LCL = 0. In addition, we note that the product based on this control chart cannot satisfy the LCL, either. Noting that $9\bar{X}$ has the gamma distribution with parameters 1 and 9 for the respective scale and shape, we have that

$$\Pr\left\{0 < \bar{X} < 2\right\} = 0.9929.$$

The portion of the acceptability in quality of the product in the process based on the above control limits would be 99.3% when the process works normally. The purpose of choosing $3-\sigma$ for the limit controls is to maintain the process with 99.9% for the acceptable quality of the product. Thus for the exponential case, *i.e.*, $\alpha = 1$, it would be difficult to keep this in the process. If we consider to use the one-sided variance, then the first definition would produce the following LSL and USL, respectively. From Table 3, we have that

$$\text{LCL} = \mu - \frac{3\sigma_L(1)}{\sqrt{n}} = 1 - \sqrt{0.418} = 0.3535,$$
$$\text{UCL} = \mu + \frac{3\sigma_R(1)}{\sqrt{n}} = 1 + \sqrt{2} = 2.4142.$$

Then we obtain that from the gamma distribution

$$\Pr\left\{0.3535 < \bar{X} < 2.4142\right\} = 0.9938.$$

Thus about 99.4% of the product in the process would be controlled as the acceptable product in quality. This increase in the controlled portion would be negligible but we note that LSL has moved significantly from 0. This means that there may remain some rooms for LSL to modify the control chart to improve the portion for the acceptable product. Especially we note that both the control limits may also satisfy both specification limits. From the second definition of the one-sided variance, we have from Table 4

$$\text{LCL} = \mu - \frac{3\sigma_L(\mu_L)}{\sqrt{n}} = 1 - \sqrt{0.0793} = 0.7184,$$
$$\text{UCL} = \mu + \frac{3\sigma_R(\mu_R)}{\sqrt{n}} = 1 + 1 = 2.$$

In addition, we have

$$\Pr\left\{0.7184 < \bar{X} < 2\right\} = 0.7886.$$

About 78.9% of the product can be controlled as the acceptable one in the process if we use this control chart, whose control limits are obtained by the one-sided variance from the second definition. Thus for the case of the control chart, it would seem to be more appropriate to use the first definition for the one-sided variance. In passing we note that the proportion of any part(lower or upper part) from the center line in those control charts do not represent 50% of the process when the process distribution is asymmetric. In this case, one may consider to use a median as the center line and may expect to contain a more acceptable proportion in the process. With this line of argument, we propose a more modified definition for C_{pk} based on a median θ as follows:

$$C_{pk}(\theta) = \min\left\{\frac{\theta - \text{LSL}}{3\sigma_L(\theta)}, \frac{\text{USL} - \theta}{3\sigma_R(\theta)}\right\}.$$

Already Park (2011) has discussed the modified $C_{pk}(\theta)$ based on the ordinary variance σ^2. For further discussion of the inference for $C_{pk}(\theta)$ you may refer to Park (2011).

The concept for the one-sided variance may be applied to obtain the confidence intervals for the location parameters when the underlying distribution are not symmetric and any pivot is not available with the same arguments used for the control chart. However in this case, the study for the coverage probability should be followed to justify the exactness and correctness of that interval.

We have proposed two types of one-sided variance. The first type has been defined as the variance for the distribution that is symmetric around u while the second one is the variance for the truncated distribution at u. As we have seen the results from the two cases, each definition of the one-sided variance has its own peculiar point to be applied. Thus the choice for the applications should depend on the purpose of the application or the shape of the underlying distribution.

For the choice for u, it should also depend on the purpose of the applications. For example if we are interested in the mean, u must be the mean of the distribution. However, for $C_{pk}(\theta)$ introduced previously we have to choose a median for u. Especially when one considers the first definition of the one-sided variance, one may choose the mode of the underlying distribution as u for the purpose of the symmetry point of each side variance. This in turn involves the estimation of the mode when one

applies the one-sided variance to the real data. Even though the burden of the estimation of the mode, one may expect that the modified definition of C_{pk} using the one-sided variance based on the mode would bring us more stable values of C_{pk}.

A referee for this paper brought the terminology of the semi-variance (Bond and Satchell, 2002) to my attention. A version of the definition for the semi-variance can be expressed as

$$\mathrm{SV}_u = \int_{-\infty}^{u} (x-u)^2 dF(x),$$

which has similar form to our first definition of the one-sided variance. To my knowledge concerns, only the lower part for u has been defined for the semi- variance. In addition, the semi-variance have been applied to finance and geophysics with the risk theory and decision of direction not to the problem with skewed distribution.

Finally, we would like to comment about the motivation of this study. As mentioned before, the goal to introduce the one-sided variance is mainly to set up some criterion or criteria for C_{pk} to easilty assess the ability of the process. Therefore for the applications of the Statistics and Probability theory to industrial fields, the application of one-sided variance should be seriously considered and in additon a study for the theoretical aspects and properties for the definition and estimation must be followed.

Acknowledgement

The author wishes to express his sincere appreciation to the referee for bringing the definition of the semi-variance to his attention.

References

Bickel, P. J. and Doksum, K. A. (1977). *Mathematical Statistics-Basic Ideas and Selected Topics*, Holden-Day, San Francisco, California.

Bond, S. A. and Satchell, S. E. (2002). Statistical properties of sample semi-variance, *Applied Mathematical Finance*, **9**, 219–239.

Chung, K. L. (1974). *A Course in Probability Theory*, 2nd Ed. Academic Press, New York.

Park, H. I. (2011). A modified definition on the process capability index C_{pk} based on median, *Communications in Korean Statistical Society*, **18**, 527–535, 286–297.

Park, S. H. and Park, Y. H. (2005). *Statistical Quality Control*, 3rd Ed., Min-Young Press, Seoul, Korea.

Received July 4, 2012; Revised August 18, 2012; Accepted August 28, 2012